White Ian Richard (Orcid ID: 0000-0002-6718-7661)

The importance of plausible data-generating mechanisms in simulation studies: A response to "Comparing methods for handling missing covariates in meta-regression" by Lee & Beretvas (doi:10.1002/jrsm.1585)

Ian R White

MRC Clinical Trials Unit at UCL

90 High Holborn

London WC1V 6LJ

UK

ian.white@ucl.ac.uk

15/7/2022

Abstract

The paper by Lee and Beretvas (doi:10.1002/jrsm.1585) described a well-executed simulation study comparing "modern" with "ad hoc" methods for performing meta-regression when some covariates are incomplete. However, they drew practical conclusions after simulating data under a single missing data mechanism which favoured the "modern" methods, while other missing data mechanisms would have favoured the "ad hoc" methods. Broad recommendations about methods to use in practice should instead be based on simulation studies using a range of plausible data-generating mechanisms. This range must represent what is believed likely to occur in practice, and not what is convenient for statistical analysis.

Introduction

The paper by Lee and Beretvas[1] described a well-executed simulation study exploring how meta-regression should be performed when, despite systematic reviewers' best efforts, some studies have missing values for one or more covariates. As a reviewer of the paper, I considered that the paper seriously over-interpreted the findings of the simulation study. The editors accepted the paper for publication and invited me to write this commentary. I will describe my concerns about this particular paper and also draw out general issues for simulation studies, especially those exploring missing data methods.

Missing data background

A principled approach to analysis with missing data identifies a plausible assumption about the missing data, performs a primary analysis that is valid under that assumption, and then explores sensitivity of the main results to departures from that assumption[2].

Assumptions often relate to the missing data mechanism, the way in which data become missing. This may be missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). In the context of meta-regression with missing covariate values, these terms describe the probability of a missing covariate value in a particular study: under MCAR, it depends on neither the missing value nor the treatment effect in

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jrsm.1605

that study; under MAR, it does not depend on the missing value, conditional on the treatment effect in that study; but under MNAR, it depends on the missing value, conditional on the treatment effect in that study.

Many modern statistical methods, such as multiple imputation (MI), commonly assume MAR, because it is convenient and efficient for estimation. This does not however make MAR correct. The method of complete-case analysis (CCA), where studies with a missing covariate value are excluded, depends on a different assumption which does not fit into the MCAR/MAR/MNAR framework: that the probability of a missing covariate value in a particular study does not depend on the treatment effect, conditional on the missing value in that study. Note the inversion of the treatment effect and the missing covariate value in this assumption, compared with MAR.

Lee and Beretvas set out to compare "modern" statistical methods that are based on a MAR assumption (MI and full information maximum likelihood, FIML) with what they call "ad hoc" approaches (CCA and shifting-case analysis, SCA, another type of complete-case analysis). The figure summarises the possible missing data assumptions, according to whether the probability of a missing covariate depends on either or both of the covariate value and the treatment effect. The figure also shows where MI and CCA are valid analyses.

[Figure about here]

<u>The authors' simulation</u>

The authors' simulation study was performed under a single data-generating mechanism, under which the data are MAR. This means the authors explored only the top left corner of the figure. This is exactly the part of the figure where MI is likely to outperform CCA. It is therefore not surprising that the "modern" methods are shown to outperform the "ad hoc" methods. I want to discuss whether the authors were right to restrict attention to this corner. This raises the question of what missing data mechanisms are plausible.

First, the authors did not simulate under an MCAR assumption. Their reason is that "study characteristics are more likely to be missing as a function of other factors (i.e., other study designs, study discipline, or even effect sizes) than to be missing completely at random" and that "MCAR is a special case of MAR". The former is probably true, and the latter is undoubtedly true. But the authors' simulation study used only one particular MAR mechanism. Can they assert that this *particular* MAR mechanism is more plausible than MCAR? This mechanism with 20% missing data sets the probability of the covariate being missing to 10%, 40% and 80% when the treatment effect g* (measured as a standardised mean difference) is 0, 1 and 2. This is a very strong relationship, which I find *less* plausible than a MCAR assumption which has 20% missing data at all levels of g*.

Secondly, the authors did not simulate under an MNAR assumption. They said "The missing not at random (MNAR) mechanism is also not considered in the current study. Researchers have agreed that there is no single, ideal way to handle MNAR data." This argument has two problems. Firstly, it ignores the specific MNAR mechanism noted above, under which CCA is valid. Secondly, it emphasises the existence or non-existence of statistical methods instead of the plausibility of the mechanism: the methods should not drive the choice of mechanism. The correct question is whether the probability of a covariate being missing is more likely to depend on the value of that covariate or on the estimated treatment effect.

Plausible missing data mechanisms

It is clear that the results of the simulation study would have been much less favourable to the "modern" methods if a broader range of data-generating mechanisms had been used, perhaps one from each corner of the Figure.

Can we assess what missing data mechanisms are plausible in practice? Missing covariate data arise from the covariate not being recorded or from its not being reported. Both seem likely to depend on the nature of the covariate in that study: for example, in studying a population where a particular subgroup is very rare, investigators might not think of recording the prevalence of that subgroup. Dependence on the *estimated* treatment effect is hard to envisage, but dependence on the *true* treatment effect in the location studied is possible through meta-confounding: for example, a poorly conducted study might fail to report covariates and have a biased treatment effect. My view is that MNAR mechanisms and MCAR mechanisms are plausible but MAR mechanisms other than MCAR are likely to be less plausible. Similar arguments have been made in study-level data[3].

I have presented my opinions of what mechanisms are plausible, but others may have different opinions. It would be helpful to have empirical evidence of what mechanisms occurs in practice, but this is especially difficult to collect in missing data problems since the missing data mechanism is unidentifiable.

Conclusions

The authors' simulation study did not support their conclusion, "we advocate the use of MI and FIML than CCA and SCA approaches in practice". This recommendation is only relevant to data sets where a particular missing data mechanism (MAR) holds, and there is no evidence that this mechanism is common or even plausible. My belief is that more extensive simulations will support the authors' conclusion for meta-regression models with *multiple* covariates, because MI and FIML make greater use of the data and hence increase precision, but not in meta-regression models with a *single* covariate.

As a general conclusion, simulation studies need to be very careful in the data-generating mechanisms used. When a method is newly proposed, it may be reasonable to choose data-generating mechanisms to favour one's proposed method, as a proof of concept, but then one can only conclude that one's method shows promise. Authors need to study a range of plausible data-generating mechanisms before they can make a recommendation that a method should be used in practice[4,5]. This range of plausible data-generating mechanisms must represent what is believed likely to occur in practice, and not what is convenient for statistical analysis.

Figure caption

Figure. The space of missing data mechanisms, showing areas representing missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR), and areas where multiple imputation (MI) and complete-case analysis (CCA) are valid. The star indicates the missing data mechanism studied by Lee and Beretvas[1].

Funding

Ian White was supported by the Medical Research Council Programme MC_UU_00004/07.

References

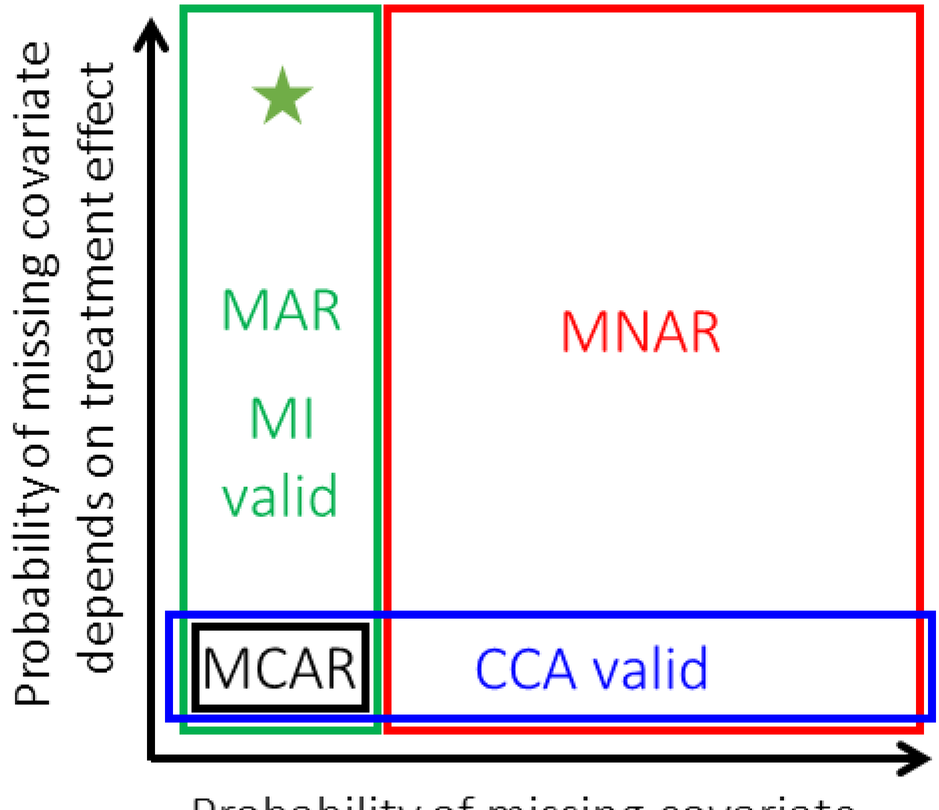
1. Lee J, Beretvas SN. Comparing Methods for Handling Missing Covariates in Meta-Regression. *Res Synth Methods*. Epub ahead of print 7 July 2022. DOI: 10.1002/jrsm.1585.

2. National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. The National Academies Press, 2010.

3. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 2010; 29: 2920–2931.

4. Boulesteix A-L, Lauer S, Eugster MJA. A Plea for Neutral Comparison Studies in Computational Sciences. *PLoS One* 2013; 8: 61562.

5. Boulesteix A-L, Binder H, Abrahamowicz M, et al. On the necessity and design of studies comparing statistical methods. *Biometrical J* 2018; 60: 216–218.


MAR
MI
valid
MNAR
MCAR
CCA valid
Probability of missing covariate depends on treatment effect
Probability of missing covariate depends on covariate value